Lecture 5: Linear regression

Statistical Learning (BST 263)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Probabilistic model for linear regression

Basis functions

Maximum likelihood estimation

Uncertainty quantification

- Distribution of $\hat{\beta}$
- Distribution of $\hat{f}(x_0)$
- Distribution of the residuals

Linear regression

- The most important statistical learning method!
- You are already very familiar with linear regression...
  - running it on data,
  - interpreting the results,
  - applying it to examples,
  - and possibly estimation.
  - (See ISL Chapter 3 for this kind of stuff.)
- So we will not rehash this stuff.
- Instead, we will do a more advanced treatment of the math behind linear regression.
- Why? It is the foundation for many, many other methods.

Outline

Probabilistic model for linear regression

Basis functions

Maximum likelihood estimation

Uncertainty quantification
- Distribution of $\hat{\beta}$
- Distribution of $\hat{f}(x_0)$
- Distribution of the residuals

Probabilistic model for linear regression

- Linear regression corresponds to using a probabilistic model based on the normal distribution.

- Training data: $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i \in \mathbb{R}$ and x_i can be in any arbitrary space.

- x_i is mapped to $\varphi(x_i) = (\varphi_1(x_i), \ldots, \varphi_p(x_i))^{\mathtt{T}} \in \mathbb{R}^p$.
- $\varphi_1, \ldots, \varphi_p$ are called the *basis functions* or *feature functions*.

What is an example of basis functions you have used before?

- The outcome y_i is modeled as a random variable

$$Y_i = \varphi(x_i)^{\mathtt{T}}\beta + \varepsilon_i$$

where $\beta \in \mathbb{R}^p$, and $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ independently.

Is the model linear in the x's, in β, or in both?

Model for linear regression – Linear algebra version

- We can describe the model more succinctly by defining $Y = (Y_1, \ldots, Y_n)^{\mathtt{T}}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathtt{T}}$, and

$$A = \begin{bmatrix} \varphi(x_1)^{\mathtt{T}} \\ \vdots \\ \varphi(x_n)^{\mathtt{T}} \end{bmatrix}.$$

What are the dimensions of A?

- Then the model is $Y = A\beta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

In words, what do Y, A, β, and ε represent?

- Equivalently, the entire model can be written in the single expression $Y \sim \mathcal{N}(A\beta, \sigma^2 I)$.

Why is this equivalent to the previous line?

Model for linear regression – Linear algebra version

- Model: $Y \sim \mathcal{N}(A\beta, \sigma^2 I)$ where $A = \begin{bmatrix}\varphi(x_1) & \cdots & \varphi(x_n)\end{bmatrix}^{\mathtt{T}}$.

- So, the density of Y (given β, σ^2, x) is

$$\begin{aligned} p(y \mid \beta, \sigma^2, x) &= \mathcal{N}(y \mid A\beta, \sigma^2 I) \\ &= \frac{1}{(2\pi)^{n/2} |\det(\sigma^2 I)|^{1/2}} \exp\Big(-\tfrac{1}{2}(y - A\beta)^{\mathtt{T}} (\sigma^2 I)^{-1} (y - A\beta) \Big). \end{aligned}$$

  Here, $x = (x_1, \ldots, x_n)$ for notational simplicity.

- $|\det(\sigma^2 I)|^{1/2} = |(\sigma^2)^n|^{1/2} = (\sigma^2)^{n/2} = \sigma^n$.

- *Can you simplify the* $\exp()$ *part to remove the matrix inverse?*

Outline

Probabilistic model for linear regression

Basis functions

Maximum likelihood estimation

Uncertainty quantification
- Distribution of $\hat{\beta}$
- Distribution of $\hat{f}(x_0)$
- Distribution of the residuals

Basis functions in linear regression

- A wide range of input-output relationships can be handled through the choice of basis functions $\varphi_1, \ldots, \varphi_p$.

- Can handle nonlinear relationships between x_i and y_i.
- The "linear" part of linear regression refers to linearity in β, not linearity in the x_i's.

  What equation are we referring to, here?

- Each x_i can be highly complex...

  e.g., images of varying size, time-series of varying length, natural language text, a collection of records, ...

- The basis functions conveniently transform x_i into a fixed-dimensionality vector of features $(\varphi_1(x_i), \ldots, \varphi_p(x_i))^{\mathtt{T}}$.

Basis functions: Common examples

- Linear with intercept:

$$\varphi(x_i) = (1,\, x_{i1},\, \ldots,\, x_{id})^{\mathtt{T}}.$$

- Quadratic:

$$\varphi(x_i) = (1, x_{i1}, \ldots, x_{id},\, x_{i1}^2, \ldots, x_{id}^2,\, x_{i1}x_{i2}, \ldots, x_{i(d-1)}x_{id})^{\mathtt{T}}.$$

- Subset of selected interactions
- Higher-order polynomials
- Splines
- Radial basis functions
- Fourier basis (sines and cosines)
- Wavelets

Basis functions: Transformations

- Dummy variables for qualitative/categorical variables:
  - Binary variable, e.g.,

    $$\mathrm{I}(\text{subject } i \text{ is female}).$$

  - Categorical variable x_{ij} taking k possible values $v_1, \ldots, v_k$: transform to $k-1$ *dummy variables*,

    $$\mathrm{I}(x_{ij} = v_1),\ \ldots,\ \mathrm{I}(x_{ij} = v_{k-1}).$$

  - If x_{ij} is a categorical variable encoded as an integer, it is important to do this transformation!
    What assumption are you making if you do not transform it?

- Fractions or percentages are often transformed using $\operatorname{logit}(x) = \log(x/(1-x))$.

- Positive numbers are often transformed using $\log(x)$.

Basis functions: Controlling flexibility

- The flexibility of a linear regression model can be controlled via the choice of basis functions.

  e.g., the number of variables to use, which variables, which interactions, the number of spline knots, etc.

- However, making this choice is sometimes difficult.

  (... both computationally and statistically)

- Often, it is easier to control flexibility using regularization.

  e.g., penalized regression such as ridge regression, lasso, or elastic net, or Bayesian linear regression.

- We will return to this later in the course.

Outline

Probabilistic model for linear regression

Basis functions

Maximum likelihood estimation

Uncertainty quantification

- Distribution of $\hat{\beta}$
- Distribution of $\hat{f}(x_0)$
- Distribution of the residuals

Maximum likelihood estimation for linear regression

- As a function of the parameters β and σ^2, $p(y \mid \beta, \sigma^2, x)$ is called the *likelihood function*.

- For the moment, let's suppose σ^2 is known.

- The log-likelihood for β is

$$\log p(y \mid \beta, \sigma^2, x) = \text{const} - \frac{1}{2\sigma^2}(y - A\beta)^{\mathtt{T}}(y - A\beta),$$

  where const denotes a constant that does not depend on β.

- A common way to estimate the parameters of a probabilistic model is to maximize the log-likelihood.

Maximum likelihood estimation for linear regression

- Maximizing the log-likelihood of β is same as minimizing

$$\begin{aligned} h(\beta) &= (y - A\beta)^{\mathtt{T}}(y - A\beta) \\ &= y^{\mathtt{T}}y - 2\beta^{\mathtt{T}}A^{\mathtt{T}}y + \beta^{\mathtt{T}}A^{\mathtt{T}}A\beta. \end{aligned}$$

- To find the minimizer, set the gradient $\nabla h(\beta)$ to zero...

$$0 = \nabla h(\beta) = -2A^{\mathtt{T}}y + 2A^{\mathtt{T}}A\beta$$

and solve for β...

$$\beta = (A^{\mathtt{T}}A)^{-1}A^{\mathtt{T}}y$$

assuming $A^{\mathtt{T}}A$ is invertible.

Can you verify the preceding steps? More advanced: Can you verify that it is a minimum (not just a critical point)?

Maximum likelihood estimation for linear regression

- Thus, the maximum likelihood estimate (MLE) is

$$\boxed{\hat{\beta} = (A^{\mathtt{T}} A)^{-1} A^{\mathtt{T}} y.}$$

- The estimated prediction function is $\hat{f}(x_0) = \varphi(x_0)^{\mathtt{T}} \hat{\beta}$.

- The MLE for σ^2 turns out to be

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}(y - A\hat{\beta})^{\mathtt{T}}(y - A\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.}$$

Outline

Probabilistic model for linear regression

Basis functions

Maximum likelihood estimation

Uncertainty quantification

- Distribution of $\hat{\beta}$
- Distribution of $\hat{f}(x_0)$
- Distribution of the residuals

Uncertainty quantification

- We can quantify our uncertainty in the estimate $\hat{\beta}$, as well as in the predictions $\hat{f}(x_0)$, by considering their probability distributions under the assumed model.
- The basic idea is to view $\hat{\beta}$ as a random vector, where the randomness comes from the outcomes Y_i in the training data $(x_1, Y_1), \ldots, (x_n, Y_n)$. The inputs x_i are treated as fixed (i.e., non-random) in this type of analysis.
- Under this setup, we can analytically derive the distributions of $\hat{\beta}$, of $\hat{f}(x_0)$, and of the residuals $Y_i - \hat{Y}_i$.

Uncertainty quantification

- These distributions are used to construct:
  - confidence intervals for the coefficient estimates,
  - p-values for testing whether coefficients are equal to 0,
  - confidence intervals for the prediction function,
  - prediction intervals for future outcomes, and
  - various residual diagnostics.
- The caveat is that these distributions are only correct when the assumed linear regression model is correct.
- In practice, the model is usually incorrect, so the resulting intervals and p-values must be viewed with skepticism.

Distribution of $\hat{\beta}$

- Under the model, $Y = A\beta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.
- $Y \in \mathbb{R}^n$ is a random vector. $A \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$ are fixed.
- So $\hat{\beta}$ is a random vector (where the randomness is from Y):

$$\begin{aligned}
\hat{\beta} &= (A^\mathsf{T} A)^{-1} A^\mathsf{T} Y \\
&= (A^\mathsf{T} A)^{-1} A^\mathsf{T} (A\beta + \varepsilon) \\
&= \beta + (A^\mathsf{T} A)^{-1} A^\mathsf{T} \varepsilon \\
&\sim \mathcal{N}(\beta, \, \sigma^2 (A^\mathsf{T} A)^{-1})
\end{aligned}$$

Can you verify the preceding steps?

Distribution of $\hat{\beta}$

- Therefore, if the model is correct, then

$$\boxed{\hat{\beta} \sim \mathcal{N}(\beta,\, \sigma^2 (A^{\mathsf{T}} A)^{-1}).}$$

- If σ^2 is known, this can be used to construct confidence intervals for the coefficients β_j, e.g., $\hat{\beta}_j \pm 1.96\sqrt{\mathrm{Var}(\hat{\beta}_j)}$.

- Usually, though, σ^2 is not known, and some more math is needed to construct correct confidence intervals when using $\hat{\sigma}^2$ instead of σ^2. We won't go into these additional details here.

Distribution of $\hat{f}(x_0)$

- If the linear regression model is correct, then

  $$\hat{f}(x_0) = \varphi(x_0)^{\mathtt{T}}\hat{\beta} \sim \mathcal{N}\Big(\varphi(x_0)^{\mathtt{T}}\beta,\, \sigma^2\varphi(x_0)^{\mathtt{T}}(A^{\mathtt{T}}A)^{-1}\varphi(x_0)\Big)$$

  by the affine transformation property. *Can you see why?*

  In words, what is this formula telling us?

- If σ^2 is known, this can be used to construct confidence intervals for $f(x_0)$ and prediction intervals for a future outcome $Y_0 = f(x_0) + \varepsilon_0$.

- As before, if σ^2 is not known, then more work is needed to construct correct confidence intervals and prediction intervals when using $\hat{\sigma}^2$.

Distribution of the residuals

- The *residuals* are the differences between the observed outcomes Y_i and the fitted outcomes $\hat{Y}_i = \varphi(x_i)^\mathtt{T}\hat{\beta}$.

- Define $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\mathtt{T}$. Then

$$\hat{Y} = A\hat{\beta} = A(A^\mathtt{T}A)^{-1}A^\mathtt{T}Y = HY$$

where $H = A(A^\mathtt{T}A)^{-1}A^\mathtt{T}$ is called the *hat matrix*.

- So the vector of residuals is

$$\begin{aligned} Y - \hat{Y} &= Y - HY = (I - H)Y \\ &\sim \mathcal{N}\Big((I - H)A\beta,\, \sigma^2(I - H)(I - H)^\mathtt{T}\Big) \end{aligned}$$

by the affine transformation property, since $Y \sim \mathcal{N}(A\beta, \sigma^2 I)$.

Distribution of the residuals

- Since $HA = A$, then $(I - H)A\beta = 0$. Further, since $H = H^{\mathsf{T}}$ and $HH = H$, then $(I - H)(I - H)^{\mathsf{T}} = I - H$. Thus,

$$\boxed{Y - \hat{Y} \sim \mathcal{N}\big(0,\, \sigma^2(I - H)\big).}$$

- If σ^2 is known, then we can compute the standardized residuals $(Y_i - \hat{Y}_i)/(\sigma\sqrt{1 - H_{ii}})$, and this result implies that they are $\mathcal{N}(0, 1)$ distributed (but not independent).

- If σ^2 is unknown, then one can derive the distribution of the studentized residuals, $(Y_i - \hat{Y}_i)/(\hat{\sigma}\sqrt{1 - H_{ii}})$.

- The definition of "standardized residuals" and "studentized residuals" varies from source to source, so you may need to be careful about precisely what definition is being used.

Leverage

- The *leverage* of point i is defined as H_{ii}, the ith diagonal entry of H.
- $\hat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j$, so if H_{ii} is large then Y_i has a large influence on the fitted value $\hat{Y}_i$.
- Identifying high leverage points is a useful diagnostic for finding points that might be having excessive influence and might be causing spurious results.
- The leverages always sum to p, i.e., $\sum_{i=1}^{n} H_{ii} = p$.

  More advanced: Can you see why this is true?